A Brief Review of Phonetic Errors in Punjabi Typed Text and Bangla

Meenu Bhagat

***Abstract*— Phonetic analysis is a branch of natural language processing (NLP) which deals with how the sounds are produced when we talk and how words are related to sounds. Phonetic analysis is a challenging task in Natural Language Processing as it involves making computer to understand how sounds are produced and analyze it. In the past few decades the researchers has addressed this problem in a broader perspective focusing on different Regional languages that are spoken across the world. The major application of Phonetics is to recognize the language, process it for lexical, syntactic, semantic knowledge about the language and the challenges of speech recognition and speech synthesis. This paper focuses on the contribution of different type of Phonetic errors in Non-word Error Distribution of Punjabi Typed Text. This paper is based on the analysis done on 20000 misspelled words generated by typists. This paper also give a brief introduction of phonetic errors in Bangla language.**

***Index Terms*— Phonetic, Gurmukhi, Nonword, Cognitive.**

I. INTRODUCTION

Error can be of two types, namely, Non-word error and Real-word error. A Real word error occurs when a word is misspelled as another valid word which is not proper for the context. An instance of a real word error is `Pu`l`→ `P`l`, here the valid Punjabi word `Pu`l` is misspelled as another valid word `P`l` .If a string of characters is separated by spaces or punctuation marks it is called a Candidate string. A Candidate string is said to be valid word if it has a meaning otherwise, it is a non-word. In each case the problem is to detect the Error and suggest correct alternatives or automatically replace it with correct word. Chaudhuri and Kundu [1] have done an elaborative analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based Bangla spellchecker. Church and W.A. Gale[2] have done a Probability scoring for spelling correction. F.J. Damerau[3] worked on a technique for computer detection and correction of spelling errors in English language. Van Berkel and DeSmedt[4] had 10 Dutch subjects transcribe a tape recording of 123 Dutch surnames randomly chosen from a telephone directory. They found that 38% of the phonetically plausible spellings generated by the subjects were incorrect .Mitton [5] found that out of 44970 of all the spelling errors in his corpus of 925 student essays involved homophones.

We can distinguish three different types of non-word misspellings [6]:

Meenu Bhagat, Assistant Professor, Department of Computer Science & Engg., Punjab University SSG Regional Centre Hoshiarpur, Punjab, India

(1) Typographic Errors - the writer or the typist knows the correct spelling but simply makes a motor coordination slip

(2) Cognitive Errors - due to misconception or lack of knowledge on the part of the writer or the typist

(3) Phonetic Errors - a special class of cognitive errors in which the writer substitutes a phonetically correct but orthographically incorrect sequence of letters for the intended words.

With an increase in Natural Language applications like Spell Checker and Corrector, Optical Character Recognition, Machine Translation, Natural Language Interfaces etc. using Indian Languages, it has become necessary to study the reasons and the nature of errors that can occur in these languages and how these errors can be detected and corrected.

Gurmukhi script[7]-consists of 41 consonants called *vianjans*, 2 symbols for nasal sounds, 9 vowel symbols called *laga* or *matras*,one symbol for reduplication of sound of any consonant and three half characters.The consonants of first row `(a,A,e)` are classified as open syllabics and called vowel consonants or semi consonants or "Matra Vahak" due to their inherent property that they are never used in work without any 'Laga' or 'Vowel'. The next two consonants are classified as root class consonants. The rest of the consonants except to the last two groups namely the - "Antim" and "Naveen" group, are categorized according to their phonetic structure.There are five such categories namely the Kavarg toli, Chavarg toli, Tavarg toli and the Pavarg toli depending upon the different organs like throat, palate, mouth, tongue and lips, using which they are pronounced or from where they originate.The last but one group consisting of 5 independent consonants `(X,r,l,v,V)` is called the "Antim" group and the last group is the (`S,^,Z,z,&,L`). "Naveen" group has been introduced to accommodate the words of Persian, Sanskrit and Arabic.

II. PHONETICALLY SIMILAR CHARACTER ERROR ANALYSIS

Phonetic errors are a special class of cognitive errors in which the writer substitutes a phonetically similar but orthographically incorrect sequence of letters for the intended word. Punjabi language also contains these type of confusion characters pairs where the typist generally type the phonetically correct but wrong character.We have classified the phonetic errors into four categories :

1. Type 1 : g→G,j→J,d→D,f→F,n→x, b→B
2. Type 2 : S→s,^→K,Z→g,z→j,&→P,L→l
3. Type 3 : yy → Y , u → U , o→ O, i → I , `→ °
4. Type 4 : r→ ® ,v→ Í ,h→ H

It has been found out that 17% of the errors are due to phonetically similar character pairs in above four categories.Out of the total no. of phonetic errors 59.28% are due to the Type 2 group elements.

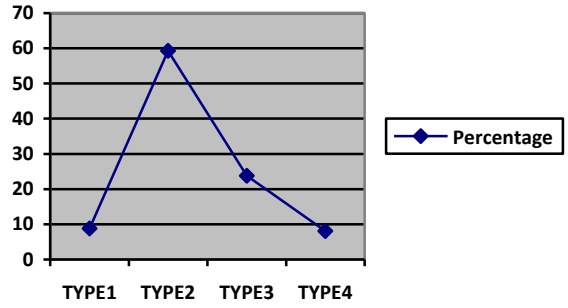

Fig1: Showing contribution of type1 to type4 Phonetic Errors

Percentage of phonetically similar sounding vowel pairs is also considerable. It has been found out that 23.83% of the misspellings contain mistakes due to Type 3 vowel pairs and 8.09% of the misspellings are due to Type 4 phonetically similar pairs.

III. Phonetic Equivalence Groups in Bangla [8]

Symbols having equivalent phonetics have been grouped together under the different Phonetic Equivalence Groups. The equivalence between symbols in a group depends on factors like the position of the symbol in a word and the matra attached to the symbol. In Bengali 135 Phonetic Equivalence Groups have been found out.An error can be generated by insertion of character matra or diacritic. Various error detection and correction procedures are written for different equivalence groups and orthographic dictionaries have also been designed on the basis of the different Phonetic Equivalence Groups.During the comparison of the input word with a word in the dictionary, if the symbols at a particular position do not match but the symbols belong to the same group then an error is signaled. The symbol in the input word is changed to the corresponding symbol in the dictionary word.

General Equivalence Groups.:

The Vowel, the Consonant and the Matra groups are always active. These are termed as the general

equivalence groups.

Special Phonetic Equivalence Groups:
Groups which are divided into some subgroups on the basis of the matra that may be attached to the symbols of the group. These subgroups are known as the Special Phonetic Equivalence Groups and have been assigned a unique Special Phonetic Equivalence Group Number.An appropriate subgroup is taken into consideration on the basis of the number set by a Phonetic Error Detection and Correction procedure. If symbols in a word belong to Equivalence Groups other than the General Equivalence Groups then more than one procedure may be executed. All the Phonetic Error Detection and Correction procedures will work on the word but it will be stored in a single dictionary.

IV. Conclusion

A detailed study has been made on the various type of phonetic errors in Punjabi Typed text. This analysis is based on the detailed analysis based on the detailed error pattern analysis that can be helpful in creating suggestion list of Punjabi spellchecker. In addition to this various other effects like phonetic effects, has also been studied. Phonetic errors can be classified into following four categories:

1. Type 1 : g→G.
2. Type 2 : S→s.
3. Type 3 : yy → Y .
4. Type 4 : r→ ®.

In addition to above a study has also been done for Bangla Text. Symbols are divided into different equivalence groups i.e. Phonetic equivalence and special Phonetic Equivalence groups .On the basis of the number set by a Phonetic Error Detection and Correction procedure, the appropriate subgroup is taken into consideration.

References

[1] P. Kundu and B.B. Chaudhuri (1999) "Error Pattern in Bangla Text". *International Journal of Dravidian Linguistics*. 28(2): 49-88.
[2] K.W. Church and W.A. Gale (1991) "Probability scoring for Spelling correction". Statistical Computing. 1(1): 93-103.
[3] F.J. Damerau (1964) "A Technique for computer detection and correction of spelling errors". *Commun. ACM.* 7(3): 171-176.
[4] VAN BERKEL, B., AND DESMEDT, K. 1988 Triphone Analysis A combined method for the correction of orthographical and typographical errors. In Proceedings of the 2nd Applied Natural Language Processing Conference (Austin,Tex., Feb.) Association for Computational Linguistics (ACL).
[5] MITTON, R. 1987. Spelling checkers, spelling correctors, and the misspellings of poor spellers. Inf, Process. Manage 23, 5, 495-505.
[6] K. Kukich (1992)"Techniques for Automatically Correcting words in Text". ACM Computing Surveys. 24(4): 377-439.
[7] Meenu Bhagat,"Difficulties in automatic text error correction in Punjabi", International Conference on Control Communication and Computer Technology" 6-7th Aug, New Delhi.